frontiers | Frontiers in Microbiology

TYPE Mini Review
PUBLISHED 05 October 2023
DOI 10.3389/fmicb.2023.1250909

OPEN ACCESS

EDITED BY
Babak Momeni,
Boston College, United States

REVIEWED BY
Sam Ma,
Chinese Academy of Sciences (CAS), China
FengLong Yang,
Fujian Medical University, China

*CORRESPONDENCE
Eliana Ibrahimi
✉ eliana.ibrahimi@fshn.edu.al
Laura Judith Marcos-Zambrano
✉ judith.marcos@imdea.org

RECEIVED 30 June 2023
ACCEPTED 22 September 2023
PUBLISHED 05 October 2023

CITATION
Ibrahimi E, Lopes MB, Dhamo X, Simeon A, Shigdel R, Hron K, Stres B, D'Elia D, Berland M and Marcos-Zambrano LJ (2023) Overview of data preprocessing for machine learning applications in human microbiome research.
Front. Microbiol. 14:1250909.
doi: 10.3389/fmicb.2023.1250909

Overview of data preprocessing for machine learning applications in human microbiome research

Eliana Ibrahimi[1]*, Marta B. Lopes[2,3], Xhilda Dhamo[4], Andrea Simeon[5], Rajesh Shigdel[6], Karel Hron[7], Blaž Stres[8,9,10,11], Domenica D'Elia[12], Magali Berland[13] and Laura Judith Marcos-Zambrano[14]*

[1]Department of Biology, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, [2]Department of Mathematics, Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal, [3]UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica, Portugal, [4]Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, [5]BioSense Institute, University of Novi Sad, Novi Sad, Serbia, [6]Department of Clinical Science, University of Bergen, Bergen, Norway, [7]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Olomouc, Czechia, [8]Department of Catalysis and Chemical Reaction Engineering, National Institute of Chemistry, Ljubljana, Slovenia, [9]Faculty of Civil and Geodetic Engineering, Institute of Sanitary Engineering, Ljubljana, Slovenia, [10]Department of Automation, Biocybernetics and Robotics, Jožef Stefan Institute, Ljubljana, Slovenia, [11]Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia, [12]Department of Biomedical Sciences, National Research Council, Institute for Biomedical Technologies, Bari, Italy, [13]INRAE, MetaGenoPolis, Université Paris-Saclay, Jouy-en-Josas, France, [14]Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, Madrid, Spain

Although metagenomic sequencing is now the preferred technique to study microbiome-host interactions, analyzing and interpreting microbiome sequencing data presents challenges primarily attributed to the statistical specificities of the data (e.g., sparse, over-dispersed, compositional, inter-variable dependency). This mini review explores preprocessing and transformation methods applied in recent human microbiome studies to address microbiome data analysis challenges. Our results indicate a limited adoption of transformation methods targeting the statistical characteristics of microbiome sequencing data. Instead, there is a prevalent usage of relative and normalization-based transformations that do not specifically account for the specific attributes of microbiome data. The information on preprocessing and transformations applied to the data before analysis was incomplete or missing in many publications, leading to reproducibility concerns, comparability issues, and questionable results. We hope this mini review will provide researchers and newcomers to the field of human microbiome research with an up-to-date point of reference for various data transformation tools and assist them in choosing the most suitable transformation method based on their research questions, objectives, and data characteristics.

KEYWORDS

human microbiome, data preprocessing, machine learning, compositionality, normalization, metagenomics data

1. Introduction

In recent decades, next-generation sequencing technologies have significantly impacted human microbiome research, allowing for a better understanding and characterization of microbiome-host interactions (Hadrich, 2020). Numerous 16S rRNA sequencing datasets are extended further by metagenomic sequencing of the whole microbial genome. The staggering increase in publications and datasets with an ever-increasing number of samples increased the need for more performant analysis approaches, such as advanced statistical methods and machine learning (ML) algorithms that can handle large-scale microbiome datasets and extract meaningful patterns, relationships, and associations. Before entering ML analysis microbiome raw data is preprocessed through several steps shown in Supplementary Figure S1.

ML models can be trained to predict the composition of microbial communities based on various input factors such as host genetics, diet, and environmental factors, which can help us understand the factors influencing microbial composition and its relation to human health (Gupta and Gupta, 2021; Hernández Medina et al., 2022). Despite the advantages, ML analysis of microbiome data is challenging due to inherent microbiome data characteristics (i.e., sparsity, compositionality, high dimensionality, dispersion), and new techniques are requested to address these challenges (Moreno-Indias et al., 2021; D'Elia et al., 2023).

Microbiome data is zero-inflated, which can be due to the sequencing depth (i.e., sampling zeros) or the real absence of taxa (i.e., true zeros) (Silverman et al., 2020). Furthermore, variations in the abundance of one taxon affect all other taxa due to the constraint that the total counts equal the library size. Hence, the raw counts observed do not directly indicate the absolute abundances of individual taxa (Weiss et al., 2017; Lloréns-Rico et al., 2021; Swift et al., 2023), giving rise to compositional data. As a result, transforming microbiome sequencing data is essential in preparing the data for analysis and applying ML algorithms.

This mini review aims to provide a comprehensive overview of the preprocessing methods used in recent human microbiome studies to transform microbiome sequencing data before ML analysis. To collect information, we conducted a scoping review based on the methodology outlined by Arksey and O'Malley (2005), combined with manual and automated literature searches following the approach outlined by Marcos-Zambrano et al. (2021). Papers included in the final review were published in peer-reviewed journals from January 2011 to January 2022 and specifically analyzed human microbiome 16S rRNA and shotgun metagenomic data through ML algorithms. As of December 2022, 3 reviewers had extracted findings on data preprocessing and transformation techniques from 95 published studies (Supplementary Table S1). In the subsequent sections, we present and discuss the findings and outcomes of our investigation.

2. Sequence preprocessing

Microbiome analysis starts with raw DNA sequencing reads or microbial taxa tables at different taxonomic resolutions, from Domain (i.e., Bacteria, Archaea, Eucarya) to strain and genome variants. Microbial taxa tables are created by processing raw sequences, known as *sequence preprocessing*. Both 16S rRNA sequencing and shotgun metagenomic sequencing generally involve preprocessing steps such as quality checking, trimming, filtering, removing, and merging (Travisany et al., 2015; Ryan et al., 2020). The key differences lie in the amplification of specific gene regions for 16S rRNA sequencing and the sequencing of entire genomes for shotgun metagenomics. The sequence preprocessing steps generally depend on the origin of the DNA sequences, sequence orientation, and sequencer type.

Quality scores are used for the recognition and removal of low-quality regions of sequence (trimming) or low-quality reads (filtration) and the determination of accurate consensus sequences (merging) (Bokulich et al., 2013). A widely adopted quality metric is the Phred quality score (Q) (Galkin et al., 2020). Then, leading, and trailing trimming are applied at the position of the read where the average score drastically changes and falls below the given threshold (Bolger et al., 2014). Typical sequence preprocessing techniques are: (1) reads filtering, if overall quality is very low (Amir et al., 2017); (2) minimal length filtering, for reads below a specified length; (3) barcode and adapter-trimming (Martin, 2011); (4) chimera filtering (Edgar et al., 2011); (5) phiX reads, commonly present in marker gene of Illumina sequence data (Callahan et al., 2016). A frequently used tool for shotgun aligning and taxonomic profiling is MetaPhlAn (Thomas et al., 2019; Blanco-Míguez et al., 2023). Shotgun metagenomics preprocessing generally requires a complex sequence of programs merged into pipelines to be used since there is no one-in-all software solution yet. The solution is usually found in automated pre-defined bioBakery Workflows (Beghini et al., 2021) or Bbtools, namely, BBMerge and BBDuk (Bushnell et al., 2017; Galkin et al., 2020).

Before entering the feature selection step, additional filtering is performed on the raw data to reduce noise while keeping the most relevant taxa. In this step, microbiome low abundance features (e.g., <500 reads) and/or prevalence (e.g., <10%) per sample group or in the entire sample, are filtered out. Based on the resulting count matrix, the taxonomic level under consideration (i.e., family, genus, species) can be chosen at this stage, considering that going down to the species level would lead to strong zero inflation.

Feature selection is approached by many studies through predictive feature selection strategies that encompass statistical methods for assessing the significance of the associations between the microbiome features and the disease condition. These methods include univariate and multivariate statistical methods, and different ML algorithms (Chen et al., 2021; Jiang et al., 2022). Network-based methods have also been employed for selecting hub strains from co-occurrence networks before entering the ML task (Xu et al., 2021). It is crucial to keep in mind that when using these predictive feature selection methods, if the training dataset is not kept distinct from the test dataset throughout all preprocessing, modeling, and assessment phases, the model gains access to test set information prior to performance evaluation, resulting in data leakage (Kapoor and Narayanan, 2022). The most common ML solution for this problem is applying a cross-validation procedure, where the initial dataset is split into several folds, and in each split, different folds are proclaimed as learning or testing folds.

3. Transformation techniques

Typically, the ML analysis of microbiome data is performed after transformations are applied to raw reads to address statistical challenges mainly associated with sparsity and the proportional nature of the generated sequencing data (Lloréns-Rico et al., 2021). Based on our review, the most common data transformation methods applied in recent human microbiome studies, in both 16 s RNA sequences and shotgun data, are the relative and normalization-based methods followed by compositional transformations such as Centered log-ratio (CLR), and Isometric log-ratio (ILR). Many reviewed publications (i.e., 28%) lack sufficient details about the data preprocessing techniques that have been applied or fail to mention if any preprocessing has been carried out leading to reproducibility issues and questionable results. In Figure 1, we present a TreeMap chart illustrating the frequencies of transformation methods applied across the analyzed papers.

Within the reviewed studies, a subset dedicated to problems of disease diagnosis and risk prediction (Fabijanić and Vlahoviček, 2016; Wu et al., 2020; Ruuskanen et al., 2021; Liu et al., 2022). Data analyzed in these studies, 16S rRNA sequencing data and shotgun data, are transformed through relative abundance, log transformations, z-score normalization, and CLR. In the following subsections, we briefly discuss the normalization-based and compositional methods applied to microbiome data before ML analysis across the reviewed papers.

3.1. Normalization methods

Two predominant transformation methods applied to deal with uneven library sizes in sequencing microbiome data are relative abundance (Statnikov et al., 2013; Ning and Beiko, 2015; Wu et al., 2018, 2021; Bogart et al., 2019; Gupta et al., 2019; Lo and Marculescu, 2019; Vangay et al., 2019; Yachida et al., 2019; Fernández-Edreira et al., 2021; Lloréns-Rico et al., 2021), and rarefaction (Stämmler et al., 2016; Weiss et al., 2017; Baksi et al., 2018), used to solve the problem of different sequencing depths (Murovec et al., 2021).

Other normalization-based methods applied frequently to microbiome data in the reviewed studies are: Log transformation, preferred when the data is heavily skewed (Lahti et al., 2013; Fabijanić and Vlahoviček, 2016; Eck et al., 2017; Tap et al., 2017; Flemer et al., 2018; Wirbel et al., 2019; Hughes et al., 2020; Ryan et al., 2020; Fouladi et al., 2021; Jiang et al., 2021; Zhu et al., 2022). Total Sum Scaling (TSS) (Lê Cao et al., 2016; Lloréns-Rico et al., 2021) which divides each taxa count by the total number of counts in each individual sample; Minimum-Maximum normalization, used to retain the relationships between the original input data (Mulenga et al., 2021; Jiang et al., 2022); Z-score normalization (Wirbel et al., 2019; Jiang et al., 2021; Mulenga et al., 2021) which transforms the data with mean zero and unit variance; the Square Root that can be successfully applied to count data that follow a Poisson distribution (Liu et al., 2011; Holmes et al., 2012); Inverse-Rank normalization used to normalize signals to approximate a normal distribution after removing the quality control sample (Ni et al., 2021).

3.2. Compositional transformations

Our review reveals a noticeable rise in the utilization of ML techniques within human microbiome research over recent years, while the adoption of compositional transformations in handling microbiome data remains relatively constrained. Nevertheless, an encouraging increasing trend in the application of compositional approaches between 2016 and 2021 is observed, as visually represented in Supplementary Figure S2. The following paragraphs delve into compositional transformations that have been employed in recent human microbiome studies, while in Table 1 we provide an overview

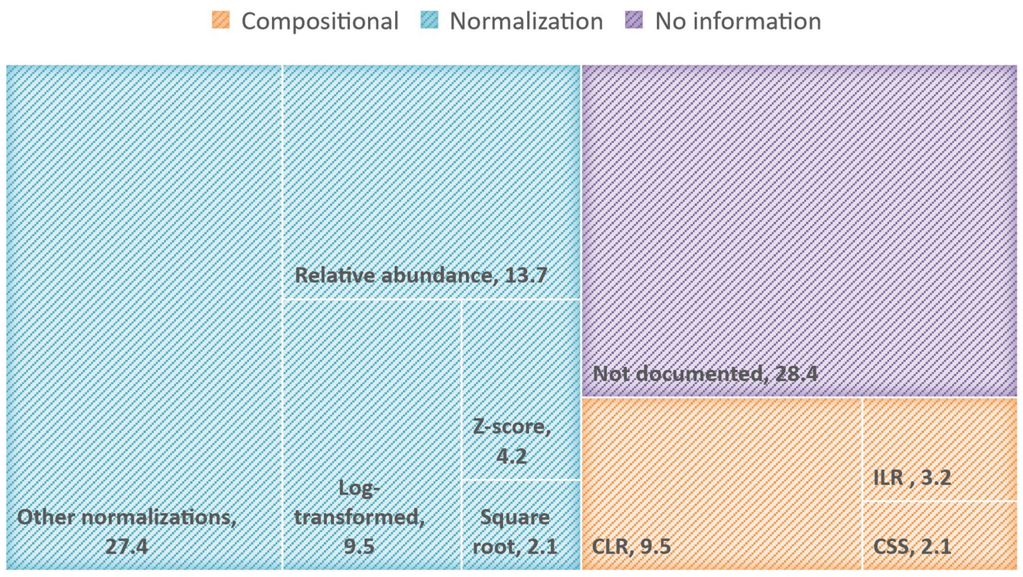

FIGURE 1
TreeMap chart illustrating the percentage of reviewed papers that applied normalization-based or compositional transformation methods, as well as the papers without clear information on preprocessing or data transformation. The other-normalization category comprises inverse-rank normalization, Box-Cox transformation, rarefaction, minimum-maximum transformation, scaling by standard deviation, normalization by total read depth, etc.

TABLE 1 Compositional transformations that are applied to human microbiome 16S rRNA and shotgun data.

Method	Bioconductor/R package	Literature
Additive log-ratio	Compositions	Aitchison (1982, 1986) and van den Boogaart and Tolosana-Delgado (2008)
Centered log-ratio	Compositions	Pawlowsky-Glahn et al. (2015) and van den Boogaart and Tolosana-Delgado (2008)
Isometric log-ratio	Compositions	Egozcue et al. (2003) and van den Boogaart and Tolosana-Delgado (2008)
Geometric mean of pairwise ratios	GMPR	Chen et al. (2018)
Trimmed mean of M-values	edgeR	Robinson et al. (2010)
Relative log expression (RLE)	edgeR	Robinson et al. (2010)
Variance-stabilizing (VST)	DESeq2	Love et al. (2014)

of the relevant literature and software tools necessary for the successful implementation of these methods.

Compositional data can be represented in a simplex space and analyzing them as absolute data with standard statistical techniques may lead to inappropriate results (Gloor et al., 2016; Quinn et al., 2018). Aitchison (1982) first proposed the additive log-ratio transformation (ALR), to address compositionality then also the centered log-ratio (CLR) (Aitchison, 1986). His followers proposed further the isometric log-ratio (ILR) (Egozcue et al., 2003; Pawlowsky-Glahn et al., 2015) and pivot log-ratio (PLR) (Filzmoser et al., 2018) transformations. The CLR transformation is applied more frequently in microbiome studies (Fabijanić and Vlahoviček, 2016; Lê Cao et al., 2016; Wirbel et al., 2019; Fukui et al., 2020; Reiman et al., 2021; Ruuskanen et al., 2021; Liu et al., 2022) than the ILR transformation (Kubinski et al., 2022), while the ALR was not applied in any of the studies included in the review.

Other compositional transformations that can be applied in microbiome data are: Cumulative Sum Scaling (CSS) (Dhungel et al., 2021; Lloréns-Rico et al., 2021), a particular representation of the relative information based on median-like quantiles; the Geometric mean of pairwise ratios (GMPR) transformation (Chen et al., 2018); the Trimmed mean of M-values (TMM) (Robinson et al., 2010); the Relative log expression (RLE) method (Robinson et al., 2010); the Variance-stabilizing transformation (VST) (Love et al., 2014).

4. Discussion

Transformations are essential for appropriately handling microbiome sequencing data, rectifying compositional issues, reducing noise, adhering to statistical assumptions, and enabling meaningful analysis and interpretation. The choice of transformation should depend on the specific characteristics of the data and the goals of the analysis. This mini review revealed substantial gaps in the process of microbiome data transformation. Relative transformations and other normalization-based methods that lead to or do not solve compositional issues (Lloréns-Rico et al., 2021) are frequently applied in recent human microbiome research.

Unlike compositional approaches (i.e., log ratios), normalization-based methods do not retrieve absolute scale from the relative data (Quinn et al., 2018). Nevertheless, when the raw data contains zero values, like in microbiome data, taking the logarithm results in negative infinity, distorting the data, and leading to invalid statistical inferences. To mitigate this issue, a pseudocount (i.e., small positive constant, ε) can be added to zero values before taking the logarithm. Selecting the right pseudocount in relation to the data's scale holds significant importance when applying log transformations (Thorsen et al., 2016). The scale of the ε, relative to the total read counts, should remain consistent across different data transformation methods applied (McKnight et al., 2019) and should be based on the context of the research problem and the scale of the data because the choice of ε can affect the results (Costea et al., 2014). Thus, it is essential to be mindful of the trade-offs between numerical stability and introducing additional bias due to the choice of ε.

Compositional transformations, ALR, CLR, and ILR log-ratio transformations, have different properties. The ALR transformation does not preserve distances because it is not isometric (Egozcue and Pawlowsky-Glahn, 2005), while CLR transformation keeps the distance, but the covariance and correlation matrix are singular because of the zero-sum of the transformed vectors (Quinn et al., 2018). In addition, aggregation of all components into the geometric mean can, in general, lead to the occurrence of false positives (Filzmoser and Walczak, 2014), so identifying the original components with the corresponding CLR variables has some limitations, which could possibly be overcome by a proper weighting strategy (Štefelová et al., 2021). Recent studies suggest that for high-dimensional compositional data, the ALR transformation should be a preferred choice for transforming variables because the interpretation of ALRs is easier than the ILR and CLR transformations (Greenacre et al., 2021). Besides log ratios, other transformations such as VST and ranked-based methods have been reported to successfully address microbiome data statistical specificities (Jeganathan and Holmes, 2021; Lloréns-Rico et al., 2021). When working with spatial human microbiome data, which can reflect the microbial composition and abundance within specific locations in the body (Adade et al., 2021), transformations for compositional spatial data that would improve ML techniques' performance when dealing with this data can be considered. Greenacre (2010, 2011) explored a power transformation that converges toward the Aitchison log-ratio transformation when the power parameter becomes 0, while Clarotto et al. (2022) propose the Isometric α-transformation (α-IT), which, unlike the ILR transformation, can successfully deal with zeros in the data.

Kubinski et al. (2022) investigated the impact of various transformation techniques on the model's predictive performance using gut microbiome data and highlighted the need to transform 16S rRNA data using compositional transformation techniques. Among the available options, the CLR transformation was identified as the most suitable, as it enables the assessment of each feature's importance in the decision-making process of ML models. Another study by McKnight et al. (2019) examined the impact of log transformations commonly employed in normalization procedures. The authors demonstrated that log transformations could distort community comparisons by suppressing significant differences in common taxa while amplifying subtle differences in rare taxa.

Thus, despite the advantages, log-ratio approaches have their limitations and drawbacks and are not the only way to deal with compositionality. Quantitative transformations such as Quantitative Microbiota Profiling (QMP) (Vandeputte et al., 2017) and Absolute Counts Scaling (ACS) (Props et al., 2017; Jian et al., 2020) offer experimental approaches to address microbiome data proportional nature. QMP involves rarefying samples to achieve an even sampling depth and scaling them based on estimated microbial loads. On the other hand, ACS directly scales the relative sequencing counts using estimated microbial loads. Lloréns-Rico et al. (2021) investigated the impact of computational and experimental techniques in addressing the issues arising from microbiome data features (i.e., compositionality and sparsity). They concluded that quantitative approaches outperform computational methods in addressing compositionality and sparsity. Authors claim that the quantitative approaches improve the identification of true positive associations while reducing the occurrence of false positives. The same study reports that when adopting quantitative methods is not feasible, computational methods that address compositionality perform better than relative methods. There are other examples in the literature where compositional methods are employed to transform microbiome data where the reader can find more details (Quinn and Erb, 2020; Yang and Zou, 2020; Greenacre et al., 2021; Yang et al., 2021; Papoutsoglou et al., 2023).

It is important to mention that in many cases the analysis of microbiome data can be performed on raw read counts rather than in transformed data. Zero-inflated negative binomial and Dirichlet-multinomial models can fit microbiome raw data quite well (Xia et al., 2018). For example, Zhang et al. (2017) applied on raw read counts a negative binomial mixed model that enables the identification of connections between the host, environmental variables, and the microbiome.

Finally, the lack of adequate information on data preprocessing and high reporting heterogeneity among papers highlight the need for standardized reporting guidelines, as also suggested by Mirzayi et al. (2021), where recommendations and guidelines are provided to help microbiome researchers properly report their findings through the 'Strengthening The Organization and Reporting of Microbiome Studies' (STORMS), composed of a 17-item checklist each related with the typical sections of a scientific paper. The omission of preprocessing and transformations applied to the data can have several significant consequences such as reproducibility concerns, misinterpretation, comparability issues, and questionable results. To mitigate these consequences, it is essential for researchers to provide thorough documentation of their data preprocessing procedures in publications. Researchers should also consider sharing their code, scripts, or workflows used for data preprocessing, which can greatly enhance transparency and reproducibility.

5. Conclusions and final remarks

Our short review shows that the utilization of data transformations that address the proportional nature of microbiome sequencing data in human microbiome studies remains limited, with many researchers primarily opting for relative and normalization-based methods that do not specifically address microbiome data characteristics. There is a lack of transparency and clear explanations regarding data preprocessing and the choice of transformation methods among the reviewed papers while it is crucial to adhere to best practices and provide a detailed methodology for developing machine learning pipelines, particularly regarding data preprocessing.

This mini review does not intend to provide unequivocal recommendations in favor of one approach over another, instead, we encourage researchers to consider the characteristics of their data carefully and whether a particular transformation method is suitable for addressing their research questions and data characteristics.

Author contributions

EI: conceptualization, investigation, writing the draft and the final manuscript. ML: investigation and writing the draft and final manuscript. XD and AS: writing the draft manuscript. RS: investigation. KH, BS, DD'E, and MB revised the draft manuscript, provided comments and writing the final manuscript. LM-Z: conceptualization, investigation, and writing the draft and final manuscript. All authors contributed to the article and approved the submitted version.

Funding

This article is based upon work from COST Action ML4Microbiome "Statistical and machine learning techniques in human microbiome studies" (CA18131), supported by COST (European Cooperation in Science and Technology), www.cost.eu. ML acknowledges support by FCT - Fundação para a Ciência e a Tecnologia, I.P., with references UIDB/00297/2020 and UIDP/00297/2020 (NOVA Math), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and CEECINST/00042/2021. KH acknowledges support through the HiTEc Cost Action CA21163 and the project PID2021-123833OB-I00 provided by the Spanish Ministry of Science and Innovation (MCIN/AEI/10:13039/501100011033) and ERDF A way of making Europe. MB acknowledges support through the Metagenopolis grant ANR-11-DPBS-0001. LM-Z is supported by Juan de la Cierva Grant (IJC2019-042188-I) from the Spanish State Research Agency of the Spanish Ministerio de Ciencia e Innovación y Ministerio de Universidades.

Acknowledgments

The authors are grateful to all the ML4Micorbiome members for the discussions and comments on this work during the ML4Microbiome meetings.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1250909/full#supplementary-material

References

Adade, E. E., Al Lakhen, K., Lemus, A. A., and Valm, A. M. (2021). Recent progress in analyzing the spatial structure of the human microbiome: Distinguishing biogeography and architecture in the oral and gut communities. *Curr. Opin. Endocr. Metab. Res.* 18, 275–283. doi: 10.1016/j.coemr.2021.04.005

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J R Stat Soc Series B.* 44, 139–177.

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems* 2:e00191-16. doi: 10.1128/mSystems.00191-16

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Baksi, K. D., Kuntal, B. K., and Mande, S. S. (2018). 'TIME': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front. Microbiol.* 9:36. doi: 10.3389/fmicb.2018.00036

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *elife* 10:e65088. doi: 10.7554/eLife.65088

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* 1–12. doi: 10.1038/s41587-023-01688-w

Bogart, E., Creswell, R., and Gerber, G. K. (2019). MITRE: inferring features from microbiota time-series data linked to host status. *Genome Biol.* 20:186. doi: 10.1186/s13059-019-1788-y

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* 12:e0185056. doi: 10.1371/journal.pone.0185056

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600

Chen, Y., Wu, T., Lu, W., Yuan, W., Pan, M., Lee, Y.-K., et al. (2021). Predicting the role of the human gut microbiome in constipation using machine-learning methods: a meta-analysis. *Microorganisms* 9:2149. doi: 10.3390/microorganisms9102149

Clarotto, L., Allard, D., and Menafoglio, A. (2022). A new class of α-transformations for the spatial analysis of compositional data. *Spat. Stat.* 47:100570. doi: 10.1016/j.spasta.2021.100570

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nat. Methods* 11:359. doi: 10.1038/nmeth.2897

D'Elia, D., Truu, J., Lahti, L., Berland, M., Papoutsoglou, G., Ceci, M., et al. (2023). Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. *Front. Microbiol.* 14:1257002. doi: 10.3389/fmicb.2023.1257002

Dhungel, E., Mreyoud, Y., Gwak, H.-J., Rajeh, A., Rho, M., and Ahn, T.-H. (2021). MegaR: an interactive R package for rapid sample classification and phenotype prediction using metagenome profiles and machine learning. *BMC Bioinformatics* 22:25. doi: 10.1186/s12859-020-03933-4

Eck, A., Zintgraf, L. M., de Groot, E. F. J., de Meij, T. G. J., Cohen, T. S., Savelkoul, P. H. M., et al. (2017). Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC Bioinformatics* 18:441. doi: 10.1186/s12859-017-1843-1

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Egozcue, J. J., and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Math. Geol.* 37, 795–828. doi: 10.1007/s11004-005-7381-9

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi: 10.1023/A:1023818214614

Fabijanić, M., and Vlahoviček, K. (2016). Big data, evolution, and metagenomes: predicting disease from gut microbiota codon usage profiles. *Methods Mol. Biol.* 1415, 509–531. doi: 10.1007/978-1-4939-3572-7_26

Fernández-Edreira, D., Liñares-Blanco, J., and Fernandez-Lozano, C. (2021). Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes. *Expert Syst. Appl.* 185:115648. doi: 10.1016/j.eswa.2021.115648

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied compositional data analysis*. Cham: Springer International Publishing.

Filzmoser, P., and Walczak, B. (2014). What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatogr. A* 1362, 194–205. doi: 10.1016/j.chroma.2014.08.050

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., et al. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67, 1454–1463. doi: 10.1136/gutjnl-2017-314814

Fouladi, F., Carroll, I. M., Sharpton, T. J., Bulik-Sullivan, E., Heinberg, L., Steffen, K. J., et al. (2021). A microbial signature following bariatric surgery is robustly consistent across multiple cohorts. *Gut Microbes* 13:1930872. doi: 10.1080/19490976.2021.1930872

Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., et al. (2020). Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome. *J. Clin. Med.* 9:2403. doi: 10.3390/jcm9082403

Galkin, F., Mamoshina, P., Aliper, A., Putin, E., Moskalev, V., Gladyshev, V. N., et al. (2020). Human gut microbiome aging clock based on taxonomic profiling and deep learning. *IScience* 23:101199. doi: 10.1016/j.isci.2020.101199

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* 26, 322–329. doi: 10.1016/j.annepidem.2016.03.003

Greenacre, M. (2010). Log-ratio analysis is a limiting case of correspondence analysis. *Math. Geosci.* 42, 129–134. doi: 10.1007/s11004-008-9212-2

Greenacre, M. (2011). Measuring subcompositional incoherence. *Math. Geosci.* 43, 681–693. doi: 10.1007/s11004-011-9338-5

Greenacre, M., Martínez-Álvaro, M., and Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. *Front. Microbiol.* 12:727398. doi: 10.3389/fmicb.2021.727398

Gupta, A., Dhakan, D. B., Maji, A., Saxena, R., P K, V. P., Mahajan, S., et al. (2019). Association of *Flavonifractor plautii*, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *MSystems* 4:e00438-19. doi: 10.1128/mSystems.00438-19

Gupta, M. M., and Gupta, A. (2021). Survey of artificial intelligence approaches in the study of anthropogenic impacts on symbiotic organisms – a holistic view. *Symbiosis* 84, 271–283. doi: 10.1007/s13199-021-00778-0

Hadrich, D. (2020). New EU projects delivering human microbiome applications. *Fut. Sci. OA* 6:FSO474. doi: 10.2144/fsoa-2020-0028

Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2:98. doi: 10.1038/s43705-022-00182-9

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One* 7:e30126. doi: 10.1371/journal.pone.0030126

Hughes, D. A., Bacigalupe, R., Wang, J., Rühlemann, M. C., Tito, R. Y., Falony, G., et al. (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* 5, 1079–1087. doi: 10.1038/s41564-020-0743-8

Jeganathan, P., and Holmes, S. P. (2021). A statistical perspective on the challenges in molecular microbial biology. *J. Agric. Biol. Environ. Stat.* 26, 131–160. doi: 10.1007/s13253-021-00447-1

Jian, C., Luukkonen, P., Yki-Järvinen, H., Salonen, A., and Korpela, K. (2020). Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One* 15:e0227285. doi: 10.1371/journal.pone.0227285

Jiang, Z., Li, J., Kong, N., Kim, J.-H., Kim, B.-S., Lee, M.-J., et al. (2022). Accurate diagnosis of atopic dermatitis by combining transcriptome and microbiota data with supervised machine learning. *Sci. Rep.* 12:290. doi: 10.1038/s41598-021-04373-7

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2021). A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* 22, 522–540. doi: 10.1093/biostatistics/kxz050

Kapoor, S., and Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. Available at: http://arxiv.org/abs/2207.07048.

Kubinski, R., Djamen-Kepaou, J.-Y., Zhanabaev, T., Hernandez-Garcia, A., Bauer, S., Hildebrand, F., et al. (2022). Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease. *Front. Genet.* 13:784397. doi: 10.3389/fgene.2022.784397

Lahti, L., Salonen, A., Kekkonen, R. A., Salojärvi, J., Jalanka-Tuovinen, J., Palva, A., et al. (2013). Associations between the human intestinal microbiota, *Lactobacillus rhamnosus* GG and serum lipids indicated by integrated analysis of high-throughput profiling data. *PeerJ* 1:e32. doi: 10.7717/peerj.32

Lê Cao, K.-A., Costello, M.-E., Lakis, V. A., Bartolo, F., Chua, X.-Y., Brazeilles, R., et al. (2016). MixMC: A multivariate statistical framework to gain insight into microbial communities. *PLoS One* 11:e0160169. doi: 10.1371/journal.pone.0160169

Liu, W., Fang, X., Zhou, Y., Dou, L., and Dou, T. (2022). Machine learning-based investigation of the relationship between gut microbiome and obesity status. *Microbes Infect.* 24:104892. doi: 10.1016/j.micinf.2021.104892

Liu, Z., Hsiao, W., Cantarel, B. L., Drábek, E. F., and Fraser-Liggett, C. (2011). Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* 27, 3242–3249. doi: 10.1093/bioinformatics/btr547

Liu, Y., Méric, G., Havulinna, A. S., Teo, S. M., Åberg, F., Ruuskanen, M., et al. (2022). Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. *Cell Metab.* 34, 719–730.e4. doi: 10.1016/j.cmet.2022.03.002

Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G., and Raes, J. (2021). Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* 12:3562. doi: 10.1038/s41467-021-23821-6

Lo, C., and Marculescu, R. (2019). MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinformatics* 20:314. doi: 10.1186/s12859-019-2833-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17:10. doi: 10.14806/ej.17.1.200

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., and Zenger, K. R. (2019). Methods for normalizing microbiome data: An ecological perspective. *Methods Ecol. Evol.* 10, 389–400. doi: 10.1111/2041-210X.13115

Mirzayi, C., Renson, A., Furlanello, C., Sansone, S.-A., Zohra, F., Elsafoury, S., et al. (2021). Reporting guidelines for human microbiome research: the STORMS checklist. *Nat. Med.* 27, 1885–1892. doi: 10.1038/s41591-021-01552-x

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Mulenga, M., Abdul Kareem, S., Qalid Md Sabri, A., Seera, M., Govind, S., Samudi, C., et al. (2021). Feature extension of gut microbiome data for deep neural network-based colorectal cancer classification. *IEEE Access* 9, 23565–23578. doi: 10.1109/ACCESS.2021.3050838

Murovec, B., Deutsch, L., and Stres, B. (2021). General unified microbiome profiling pipeline (GUMPP) for large scale, streamlined and reproducible analysis of bacterial 16S rRNA data to predicted microbial metagenomes, enzymatic reactions and metabolic pathways. *Metabolites* 11:336. doi: 10.3390/metabo11060336

Ni, Y., Lohinai, Z., Heshiki, Y., Dome, B., Moldvay, J., Dulka, E., et al. (2021). Distinct composition and metabolic functions of human gut microbiota are associated with cachexia in lung cancer patients. *ISME J.* 15, 3207–3220. doi: 10.1038/s41396-021-00998-8

Ning, J., and Beiko, R. G. (2015). Phylogenetic approaches to microbial community classification. *Microbiome* 3:47. doi: 10.1186/s40168-015-0114-5

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol.* 14:1261889. doi: 10.3389/fmicb.2023.1261889

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modelling and analysis of compositional data*. Chichester: John Wiley & Sons, Ltd.

Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman, W., et al. (2017). Absolute quantification of microbial taxon abundances. *ISME J.* 11, 584–587. doi: 10.1038/ismej.2016.117

Quinn, T. P., and Erb, I. (2020). Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. *MSystems* 5:e00230-19. doi: 10.1128/mSystems.00230-19

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34, 2870–2878. doi: 10.1093/bioinformatics/bty175

Reiman, D., Layden, B. T., and Dai, Y. (2021). MiMeNet: Exploring microbiome-metabolome relationships using neural networks. *PLoS Comput. Biol.* 17:e1009021. doi: 10.1371/journal.pcbi.1009021

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Ruuskanen, M. O., Åberg, F., Männistö, V., Havulinna, A. S., Méric, G., Liu, Y., et al. (2021). Links between gut microbiome composition and fatty liver disease in a large population sample. *Gut Microbes* 13, 1–22. doi: 10.1080/19490976.2021.1888673

Ryan, F. J., Ahern, A. M., Fitzgerald, R. S., Laserna-Mendieta, E. J., Power, E. M., Clooney, A. G., et al. (2020). Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* 11:1512. doi: 10.1038/s41467-020-15342-5

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., et al. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 4:28. doi: 10.1186/s40168-016-0175-0

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1:11. doi: 10.1186/2049-2618-1-11

Štefelová, N., Palarea-Albaladejo, J., and Hron, K. (2021). Weighted pivot coordinates for partial least squares-based marker discovery in high-throughput compositional data. *Stat. Anal. Data Mining ASA Data Sci. J.* 14, 315–330. doi: 10.1002/sam.11514

Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., and Wei, X. (2023). A review of normalization and differential abundance methods for microbiome counts data. WIREs. *Comput. Stat.* 15:e1586. doi: 10.1002/wics.1586

Tap, J., Derrien, M., Törnblom, H., Brazeilles, R., Cools-Portier, S., Doré, J., et al. (2017). Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. *Gastroenterology* 152, 111–123.e8. doi: 10.1053/j.gastro.2016.09.049

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., et al. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4:62. doi: 10.1186/s40168-016-0208-8

Travisany, D., Galarce, D., Maass, A., and Assar, R. (2015). "Predicting the metagenomics content with multiple CART trees" in *Mathematical Models in Biology* (Cham: Springer International Publishing), 145–160.

van den Boogaart, K. G., and Tolosana-Delgado, R. (2008). "compositions": A unified R package to analyze compositional data. *Comput. Geosci.* 34, 320–338. doi: 10.1016/j.cageo.2006.11.017

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551, 507–511. doi: 10.1038/nature24460

Vangay, P., Hillmann, B. M., and Knights, D. (2019). Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. *GigaScience* 8:giz042. doi: 10.1093/gigascience/giz042

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., et al. (2018). Metagenomics biomarkers selected for prediction of three different diseases in Chinese population. *Biomed. Res. Int.* 2018, 1–7. doi: 10.1155/2018/2936257

Wu, S., Chen, Y., Li, Z., Li, J., Zhao, F., and Su, X. (2021). Towards multi-label classification: Next step of machine learning for microbiome research. *Comput. Struct. Biotechnol. J.* 19, 2742–2749. doi: 10.1016/j.csbj.2021.04.054

Wu, T., Wang, H., Lu, W., Zhai, Q., Zhang, Q., Yuan, W., et al. (2020). Potential of gut microbiome for detection of autism spectrum disorder. *Microb. Pathog.* 149:104568. doi: 10.1016/j.micpath.2020.104568

Xia, Y., Sun, J., and Chen, D.-G. (2018). *Statistical Analysis of Microbiome Data with R*. Springer: Singapore.

Xu, C., Zhou, M., Xie, Z., Li, M., Zhu, X., and Zhu, H. (2021). LightCUD: a program for diagnosing IBD based on human gut microbiome data. *BioData Mining* 14:2. doi: 10.1186/s13040-021-00241-2

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7

Yang, F., and Zou, Q. (2020). mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. *Database* 2020:baaa050. doi: 10.1093/database/baaa050

Yang, F., Zou, Q., and Gao, B. (2021). GutBalance: a server for the human gut microbiome-based disease prediction and biomarker discovery with compositionality addressed. *Brief. Bioinform.* 22:bbaa436. doi: 10.1093/bib/bbaa436

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., et al. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics* 18:4. doi: 10.1186/s12859-016-1441-7

Zhu, C., Wang, X., Li, J., Jiang, R., Chen, H., Chen, T., et al. (2022). Determine independent gut microbiota-diseases association by eliminating the effects of human lifestyle factors. *BMC Microbiol.* 22:4. doi: 10.1186/s12866-021-02414-9